Chapter 8

Summarizing discussion

ABSTRACT

Genetical genomics integrates data from multiple molecular levels such as the transcriptome, proteome and metabolome by mapping their variation in a population to polymorphic genetic loci. This systems genetics approach is increasingly used to identify molecular traits involved in the pathology of diseases and to elucidate the networks underlying complex phenotypes. Recent studies have pushed the genetical genomics concept further towards data integration and interpretation within and across molecular levels, and have also revealed remaining challenges. The focus of this chapter is to discuss these challenges and their possible solutions in the following three following areas: (1) experimental design, (2) setting significance thresholds, and (3) defining gene and QTL networks. Finally, we explore how future genetical genomics studies might benefit from the advent of new methods that aim at removing large pervasive variation components that are caused by uncontrolled factors in omics datasets.

8.1 Introduction

Genetical genomics (Jansen 2003, Jansen and Nap 2001) uses classical genetics approaches of Quantitative Trait Locus (QTL) mapping to link or associate the variation in traits from multiple molecular levels (such as transcriptomics, proteomics and metabolomics) to genetic loci harboring genotypic polymorphisms. Genetical genomics has become a popular systems genetics strategy (Sieberts and Schadt 2007) for unraveling molecular regulatory networks: a PubMed search on relevant keywords currently yields 191 scientific publications (webCite 2010), 39% of which were published in 2009/10. Pioneering experiments have demonstrated the high heritability of an extensive range of molecular traits (mainly mRNAs but also protein and metabolite abundance as measured with mass spectrometry or nuclear magnetic resonance, see Box 8.1 for a summary of special features of molecular traits) in numerous model species (including yeasts, plants, worms, flies, mice, rats and humans) (Brem et al. 2002, Bystrykh et al. 2005, Chesler et al. 2005, DeCook
et al. 2006, Petretto et al. 2006, Ruden et al. 2009, Schadt et al. 2003), and they have exposed the plasticity of the QTLs that control those traits with respect to environmental condition, tissue type or cellular context (Dimas et al. 2009, Ge et al. 2009, Gerrits et al. 2009, Li et al. 2006b, Smith and Kruglyak 2008). Genetical genomics studies that integrate ‘classical’ phenotypes (such as height or disease susceptibility) with multiple traits from molecular levels have improved our understanding of how genetic variation propagates through biological systems (Fu et al. 2009) and have suggested molecular pathways through which some genetic variants can cause diseases (Emilsson et al. 2008, Moffatt et al. 2007, Schadt et al. 2008). While scaling up classical quantitative genetics approaches to the study of thousands of omics traits opens new avenues for the dissection of molecular mechanisms that regulate biological systems, it is also accompanied by a whole new range of specific challenges. These challenges are intrinsic to the high-throughput nature of the measurements, to technical aspects of the profiling technologies used, to the statistical issues introduced by the untargeted multifactorial perturbation that underlies the approach, and to the complexity of the molecular networks under study. In this review, we inspect important issues that arise at each step (Figure 8.1) of a genetical genomics study from experimental design to result interpretation. We provide accordingly recommendations allowing a reliable and efficient use of the power of genetical genomics. This review will focus primarily on gene expression profiling, but many of the issues raised are also applicable to other “omics” technologies.

**BOX 8.1: Special features of molecular traits**

Many types of molecular traits studied in genetical genomics experiments have specific properties that can be helpful in the analysis and interpretation of the data. For example, the location of the genes coding for the transcripts or the proteins studied is usually known. This extra information compared to classical non-molecular phenotypes can be used to gain a deeper insight into the mechanistic details of the underlying biological processes.

**Isoforms and modifications**

Molecular traits can often be observed in different forms. For example, transcripts are spliced into different variants. It was shown that this alternative splicing could itself be mapped to genetic variation (Kwan et al. 2008, Li et al. 2010a). On a 2D-gel, the same protein often migrates to more than one spot. This can be explained by post-translational modifications such as phosphorylation, and mapping the genetic basis for variations in the phosphoproteome and kinome is an exciting prospect. Similarly with mass spectrometry techniques, different forms of the same protein will yield different sets of peaks. Moreover, in order to remove the multiple testing induced by the observation of the same protein through multiple mass peaks, statistical methods that automatically connect those peaks can advantageously be used (Dijkstra and Jansen 2009).
BOX 8.1(continued): Special features of molecular traits

Local and distant QTLs
Protein and mRNA eQTLs can be classified into local and distant QTLs depending on the relative location of the QTL and the gene coding for the measured mRNA or protein. The identification of distant QTLs tends to be less reliable than that of local QTLs: this can be attributed to two main reasons. First, the effects of distant QTLs are biologically more indirect and therefore harder to pick up. Second, when one tests for a distant QTL, the loci tested are genome wide, as opposed to just the gene’s locus in the case of a local effect, therefore the power to detect distant QTLs is much lower than the power to detect local QTLs. This is consistent with the fact that distant QTLs tend to be more difficult to replicate than local QTLs (Peirce et al. 2006).

False local QTLs
Frequently, strong local linkage affecting transcripts reported intensity is actually the result of a technical artifact rather than a biological differential expression. Microarray probes are most often designed following reference assemblies that are based on “mainstream” laboratory strains (in the case of model organisms) or a few individuals (in the case of humans). As a result of this, the hybridization efficiency of these probes can vary from one individual to another when the probe’s target sequence is polymorphic (Alberts et al. 2007). In this case, the differential intensity between the two alleles will reflect a difference in hybridization efficiency rather than a difference in true gene expression signal. It is therefore necessary to mask all probes containing known sequence polymorphisms between the parental strains studied. Alternatively, for short-oligonucleotide arrays, which typically contain multiple probes per gene, Alberts et al. suggest to include probe effects in the statistical model, which can potentially separate “true” differential mRNA expression from “ghost” effects caused by polymorphisms (Alberts et al. 2005).

8.2 Designing a genetic experiment for thousands of phenotypes

Many of the considerations that apply to the experimental design of a classical genetic study also apply to genetical genomics. However, because the number of traits studied in a genetical genomics experiments is of a much higher magnitude (tens of thousands typically), a few specific issues need to be taken into account when deciding the population type, the sample size, and the assignment of samples to different treatments or conditions. In this section, we address the consequences of these decisions, and we introduce approaches for optimizing the power and resolution in genetical genomics experiments.
Figure 8.1: Genetical genomics as a systems genetics strategy. Genetical genomics integrates data from multiple molecular levels (middle box) and classical phenotypes (upper box) by relating their variation to common multifactorial perturbations (a combination of (epi)genetic and environmental factors as shown in the lower box). Methods to form undirected association networks within a molecular level (e.g. coexpression network) or to draw directed edges between phenotypes use QTL information and allow relating classical phenotypes such as obesity to relevant genes, metabolites or proteins.

8.2.1 Population

Just as in any genetic study, the first critical step in designing a genetical genomics experiment is the choice of a population to be studied, which will determine the
ensuing mapping strategy: linkage or association mapping. In genetical genomics studies, multiple testing caused by the mapping of large numbers of phenotypes reduces the available statistical power. Linkage mapping on an inbred population such as recombinant inbred lines (RILs), F2 intercrosses or backcrosses provides enough power to perform eQTL studies with relatively small sample sizes. Fully inbred populations (immortal lines) allow collecting different types of phenotypes on distinct but genetically identical individuals, which is a valuable advantage in systems biology experiment where invasive procedures are needed to collect various phenotypes. However, linkage genetical genomics studies in general provide a relatively poor resolution, i.e. the confidence intervals surrounding a QTL span large genome regions of typically several million base pairs. This results in additional efforts needed to identify the actual polymorphism causing the QTL, for instance using independent information such as gene annotation (Franke et al. 2006), or generating congenic strains. RILs can be generated from more than two parental strains; for example, the mouse collaborative cross (Churchill et al. 2004) and the Multiparent Advanced Generation Inter-Cross (MAGIC) for Arabidopsis thaliana (Kover et al. 2009) use 8 and 19 founder lines, respectively. Other types of crosses, such as Advanced Intercross Lines (AIL) (Darvasi and Soller 1995, Rockman and Kruglyak 2008) or Heterogeneous Stock (HS) for rat (Hansen and Spuhler 1984), introduce more recombinations and therefore have improved resolution. Association studies performed on natural or outbred populations on the other hand, have less power because a much larger number of smaller genomic regions are tested for QTL, leading to a drop in statistical significance caused by increased multiple testing and because of the large imbalance of the allele frequencies of genotypes. Since association studies allow for a much finer mapping of the QTL than that obtained with linkage analysis, there is a trade-off to consider between power and resolution when choosing the mapping strategy. Genome-wide association studies (GWAS) have naturally been used to perform genetical genomics studies in humans (Dubois et al. 2010, Emilsson et al. 2008, Goring et al. 2007, Heap et al. 2009, Stranger et al. 2005) and are emerging in model organisms studies using outbred populations (Ghazalpour et al. 2008).

8.2.2 Combining studies

Combining information from different studies can further increase the power and resolution in eQTL mapping. Meta-analysis of multiple datasets is a strategy widely used in GWAS of classical traits but is only starting to be explored in the context of genetical genomics (Dubois et al. 2010, Heap et al. 2009). Meta-analyses use sta-
8. Summarizing discussion

Statistical methods for combining p-values (Whitlock 2005), because combining directly data from different experiments is hampered by heterogeneity issues (e.g. different microarray platforms). As a result the power increases: combining their own peripheral blood dataset with the HapMap B-cell dataset, Heap et al. report close to 40% additional eQTLs that were not detected in the individual eQTL scans. Also, the combination of association and linkage mapping, a procedure commonly used in classical genetics studies, has recently been applied to eQTL studies (Gatti et al. 2009). A linkage study is first performed to identify eQTL regions with satisfactory power; an association study is then performed to refine the eQTL found by the linkage study. This association step can be performed using a relaxed statistical significance threshold since only the regions identified in the linkage step are tested.

8.2.3 Sample assignment for molecular profiling

After the choice of a population to be studied, molecular profiling can be conducted using high-throughput technologies. In this step, random assignment of experimental units is a fundamental principle of experimental design which ensures that a treatment of interest is not confounded with other factors (Fisher 1951, Wit and McClure 2004). While the genotypes are naturally randomized in the process of meiotic recombination and segregation, randomization must be enforced for other relevant factors during the design of genetical genomics experiments. In order to optimize the design for statistical power, the best way is to increase the sample size; but a smart assignment of samples to experimental units can further maximize the information that can be extracted from the data without any additional costs. For example, it was suggested to pair the most genetically distant individuals on two-color microarrays so as to maximize the number of informative genetic contrasts (Fu and Jansen 2006, Lam et al. 2008). Two-color arrays are no longer widely used, but the basic idea can be elegantly generalized: in genetical genomics experiments studying environmental perturbation, one aims at achieving the most accurate estimate of the QTL effects and QTL-by-environment interaction effects of interest, either in one or more regions of special interest, such as a previously detected phenotypic QTL, or across the entire genome. In this case, genotyped individuals can be ‘intelligently’ selected and distributed across multiple environments using an optimization algorithm to minimize the sum of variance of the parameter estimates of interest (Lam et al. 2008, Li et al. 2008, Li et al. 2009).
8.3 Significance thresholds for eQTL detection

The large number of molecular traits (tens of thousands) and markers (from 100s to millions) that are tested in a genetical genomics study requires the significance level for linkage or association to be rigorously adjusted to control the number of false positive results. Bonferroni correction in this context tends to be too conservative, and in genetical genomics studies, it is more appropriate to control false discovery rate (FDR) (Benjamini and Hochberg 1995). In practice, the approaches for calculating the significance thresholds accounting for multiple testing used in genetical genomics are mostly relying on permutations (Breitling et al. 2008a, Churchill and Doerge 1994), since standard approaches (Storey and Tibshirani 2003) work under the assumption that there is relatively mild dependence of the tests, which is not the case in genetical genomics where important correlations exist between traits and between neighboring markers. Permuting aims at breaking the biological relationship between genotypes and traits so that any QTL detected in the permuted dataset is a false positive, which allows estimating the FDR by providing an estimate of the number of false positives to be expected in the original data. By permuting only the sample labels in the genotype data, both the correlation structure between traits and the correlation structure between markers is conserved, which makes this empirical procedure perfectly suited to a non-biased estimation of the significance under the multiple dependences present in the data. If a major correlation structure is causing large groups of genes to be associated with the genotypes at random genomic loci, forming spurious hotspots of eQTLs, such permutations would also be likely to lead to hotspots being mapped by chance and therefore identify the hotspots as not significant (Breitling et al. 2008a). Thousands of permutations are usually required to ensure accuracy of the FDR estimates, but methods approximating the tail of the distribution may allow for extrapolation from a smaller number of permutations and reduce the computational burden (Knijnenburg et al. 2009). When the statistical models used for mapping contain genetic, environmental and interacting factors, the appropriate permutation strategy may be difficult to determine as certain situations require different permutation procedures to be used for individual terms in the ANOVA model, including restricted permutation, permutation of whole groups of units, permutation of some forms of residuals or some combination of these (Anderson and ter Braak 2003).

Special situations require some additional adjustments to the significance threshold used. Firstly, testing for a local eQTL effect (a QTL affecting a gene lying in a nearby locus on the same chromosome) involves testing the genotypes at only one restricted genome region as opposed to the whole genome when scanning for dis-
tant genetic effects. Therefore detection of local eQTLs is affected to a much lesser extent to multiple testing and it is advisable to use a relaxed threshold for the detection of local QTLs. Secondly, in the presence of imbalanced allele frequencies (occurring randomly or caused by segregation distortion) in an experimental population, one of the genotype group may have a very limited size yielding unreliable estimate of mean within that group, which in turn may influence the accuracy of the p-value estimates. The same issue is usually avoided in association studies where SNPs with very low minor allele frequency (e.g. below 5%) are simply excluded, at the risk of missing important biological phenomena (Dickson et al. 2010).

8.4 Defining gene and QTL networks

In addition to the genetic dissection of phenotypic variation using QTL mapping techniques, systems geneticists are interested in reconstructing the biological networks that connect genes, proteins and other traits based on their observed genetic (co-)variation. In this context, biological networks are often defined by graphical models that are composed of nodes representing traits such as gene expression levels and edges representing (causal, correlational or mechanistic) relationships between these nodes. In current genetical genomics studies, there are two main types of approaches for the inference of such networks (i) methods for identifying coexpression networks on the basis of (partial) correlations between traits; (ii) methods for identifying QTL networks on the basis of QTL underlying variation and coexpression.

8.4.1 Correlation-based networks

Coexpression networks are undirected networks in which edges connect genes that have correlated expression behaviors over a set of samples. In the genetical genomics context, these samples come from genetically diverse individuals, possibly observed over multiple conditions. The coexpression similarity between genes can be measured using different metrics, the most commonly used being Pearson’s correlation. Coexpression of two genes does not prove a causal relationship; however under the principle of “guilt by association”, it can be used to predict similar gene functions and is indicative of possible co-regulation.

Coexpression methods can be divided between unweighted and weighted approaches. Unweighted approaches use cutoffs to define a minimum level of correlation (or of another coexpression metric) required to draw an edge between two
8.4 Defining gene and QTL networks

genes. Those cutoffs are typically determined using permutations (Butte et al. 2000, Carter et al. 2004). Weighted approaches on the other hand consider all genes to be interconnected albeit with different strengths (weights) (Zhang and Horvath 2005). Weighted approaches typically transform the correlation (e.g. with a power function) to emphasize the weight of higher correlations, a practice known as “soft thresholding”. A natural step following the generation of the global network in a co-expression analysis is to identify modules of coexpressed genes, i.e. find communities of highly interconnected nodes within the graph. Such coexpressed modules can then be studied as putative functional units, thereby considerably reducing the dimensionality of the data. Different approaches have been proposed, many of which are inspired by social network research. Chesler et al. choose to focus on sets of genes in which all nodes are interconnected; such sets are termed “cliques” (Chesler et al. 2005). Searching for cliques in a network containing thousands of nodes poses a serious computational burden and several algorithms have been designed to alleviate it (Baldwin et al. 2005). An alternative is the use of the topological overlap measure (TOM): this metric allows grouping together genes that share the same neighbors in the correlation graph (Ravasz et al. 2002, Zhang and Horvath 2005), but without the strong constraint imposed by cliqueness.

One strategy to identify important genes within coexpressed modules has been to focus on highly connected genes. Connectivity (also known as degree) represents the amount of edges reaching a gene in the coexpression network in the unweighted case, or the sum of the correlation strength with all other genes (correlations) in the weighted case. Genes with high connectivity, termed “hubs”, have been claimed to be enriched for essential genes (Carter et al. 2004). Connectivity is therefore used to prioritize between genes belonging to modules of interest.

Similar correlation-based approaches can be used to study metabolites (Kose et al. 2001). Steuer discussed the important differences existing in the correlation structure of metabolites compared to that of genes because of the specific biochemical characteristics of metabolic networks, in which molecules rather than information is flowing along pathways (Steuer 2006). A promising perspective is the profiling of multiple classes of macromolecules in the same samples in order to form correlation networks integrating genes, metabolites, and possibly proteins (Fu et al. 2009).

By using partial correlations, i.e. conditioning on selected other nodes in the network, it is possible to remove indirect edges from the network (Bing and Hoeschele 2005, de la Fuente et al. 2004, Keurentjes et al. 2006). Since large scale changes in coexpression may indicate rewiring of the transcriptional network, recent work has
focused on the identification of such changes between different conditions in what is known as differential coexpression analysis (Choi and Kendziorski 2009, Tesson et al. 2010). One limitation of correlation-based networks is that they are undirected and do not use explicitly the genotypic variation, therefore lacking the causal information that is needed to identify the drivers of biological processes.

8.4.2 QTL-based networks

The interest of using multiple QTL co-localization information for the reconstruction of trait networks has been noted early on (Jansen and Nap 2001). The basic idea is that QTLs from upstream regulators should also be QTLs of the associated downstream traits, providing a simple means to order traits from causal to reactive. Moreover, when two genes map to the same eQTL, one locally and one distantly, the gene with the local eQTL is likely to regulate the gene with a distant eQTL (Jansen 2003). In practice, the application of these ideas has been hampered by two limitations of most available datasets. Firstly, the lack of power of current genetical genomics experiments does not allow for deconvolution of traits into multiple QTL (one or two QTL per trait are detected at best, and discrimination between a weak but existing QTL and absence of any QTL effect is difficult). Secondly, in experiments with low mapping resolution, it is often impossible to discriminate between two distinct neighbouring QTLs, and one shared QTL (statistical methods provide ‘parsimonious’ models, but this does not exclude that reality is more complex).

Building on the aforementioned fundamental principles, Bayesian modeling concepts for causal inference have been adapted to assist in the extraction of regulatory evidence from genetical genomics data. If a trait T1 regulates a trait T2, then variation in T1 will be propagated to T2. When some of T1’s variation can be accounted for by a QTL, this QTL will also explain some of the variation in T2. The regression of T2 on T1 corrects T2 for the variation propagated from T1, including the QTL variation: this independence of T2 and the QTL conditional on T1 is used as evidence for the fact that T1 is causal for (regulates) T2. Different statistical testing frameworks have been proposed to use this conditional independence property. For example, model selection approaches have been used to identify the causal relationship among traits that is best supported by the data (Li et al. 2006a, Schadt et al. 2005). Chen et al. provided a method to quantify the likelihood of each causal link (Chen et al. 2007). Recently, Millstein et al. further formalize a similar idea into a hypothesis test which results in a quantitative estimation of significance in terms of p-value (Millstein et al. 2009). Chaibub Neto et al. propose a likelihood-based method to compare graph configurations in which the non-propagated vari-
8.4 Defining gene and QTL networks

8.4.2 Non-shared QTLs

The performance of those methods in terms of power, false positive and false negative rates are strongly dependent on sample size, QTL effect sizes, genotype frequencies and measurement errors (Li et al. 2010b).

Some attempts have been made to combine co-expression networks with QTL-based causal inference: either by orienting undirected edges of coexpression networks (Aten et al. 2008, Chaibub Neto et al. 2008) or by inferring causal relationships between entire modules and clinical traits by studying the eigengenes representing those modules or selected genes from those modules (Chen et al. 2008, Plaisier et al. 2009).

8.4.3 Hotspots

A particular case of QTL-based networks is that of QTL hotspots: specific loci that control a large number of genes distantly. Hotspots may be the consequence of one single polymorphism with major direct effects: for example, a polymorphic transcription factor affecting multiple targets. Hotspots could also be the result of the indirect downstream effects of a single polymorphism. A handful of such eQTL hotspots have been biologically validated. For example, using a small interfering RNA (siRNA) knockdown of selected candidate genes, Wu et al. were able to observe a phenotypic change consistent with the predicted function of the hotspot genes in mouse (Wu et al. 2008). Also, a variant in the ERECTA gene was found to cause variation in a number of molecular traits (transcripts, proteins and metabolites) as well as classical phenotypes (Fu et al. 2009). Another possible hotspot cause is the co-localization of multiple polymorphisms with unrelated effects: the hotspot QTL would then typically lie within a SNP dense region or a gene rich region and this can be tested.

If the hotspot is the result of a single polymorphism, one might expect that genes whose expression is affected by this polymorphism should belong to a common biological pathway or process, at least if the effect is reasonably direct. For that reason, one of the first tests performed on the genes affected by a hotspot is often a gene annotation enrichment analysis such as GSEA (Backes et al. 2007, Dennis et al. 2003) or iGA (Breitling et al. 2004). It is, however, important to remember that this annotation enrichment alone does not prove the validity of the hotspots: as detailed in Box 8.2 with the tissue purity scenario, spurious hotspots would be expected to contain biologically related genes as well.

The search for a “master regulator” within the hotspot QTL interval is challenging since typically many candidate genes lie in the QTL confidence interval due
to the lack of resolution in most genetical genomics linkage studies (see also the earlier section). Interestingly, loci harboring eQTL hotspots were not found to be enriched for transcription factors in a yeast study (Yvert et al. 2003), and the majority of hotspots turns out to be due to very indirect effects on gene expression. In order to prioritize genes within the list of candidate regulators, multiple independent sources of information can be utilized (Franke et al. 2006). Statistical evidence such as correlation of the hotspot genes with the candidate regulator or the presence of a local eQTL for the regulator can be integrated with biological evidence such as the relevance of the functional annotations associated with the candidate gene. Sequence information can also be used. Is the candidate gene polymorphic between the two parental strains? Is there evidence of enrichment of certain transcription factor binding sites within the hotspot target genes that would provide clues on the involvement of a certain regulator? Finally, it is important to remember that the regulators underlying the QTL may not be protein-coding genes but could also be miRNAs, or structural or epigenetic mechanisms. For integrating these different pieces of information, the rank product method can be applied to prioritize the candidate regulators by multiplication across the ranks positions of candidate genes in each prioritization step (Breitling et al. 2004, Keurentjes et al. 2007).

Regulatory links derived from the above-mentioned approaches should only be considered as putative and should be experimentally validated. Experimental validation techniques include gene knockouts, transgenic animals, RNA interference-based knockdowns and chemical perturbations.

8.4.4 Non-genetic variation

The aim of genetical genomics studies is to dissect genetic variation and trace its propagation from DNA to molecular and classical traits shedding some light onto biological pathways and processes. However, previous studies have shown that many non-genetic factors contribute to variation in the data (Churchill 2002). For example, hybridization batches, population stratification, the preparation of samples by different technicians, variations in sample composition and purity can introduce strong expression changes in large groups of genes, creating a pervasive correlation structure. When those factors are known, there effects can be corrected (by including them in a regression), as is routinely done with controlled factors such as sex or age. Unfortunately, in many cases, the relevant factors are “hidden” (e.g. varying physiological state or tissue purity in the samples), and failure to account for them will have a crucial influence at many levels of a genetical genomics study.
8.4 Defining gene and QTL networks

BOX 8.2: Confounding factors for eQTL hotspots.

**Known confounding factors**

Many uncontrolled factors, such as hybridization batches, can result in highly correlated traits and clusters of QTLs. These non-biological sources of variation can be introduced during the preparation of samples, the hybridization and the measurements. Known batch effects can be filtered out from the data using a batch correction method, such as a linear regression. In some cases, however, the confounding factors are unknown and this strategy cannot be applied. When applying a genetical genomics strategy to multiple tissues/conditions, the scale of the experiment is usually doubled or tripled. It is almost impossible to conduct the whole experiment at once. The resulting batch differences become influential and it should be noted that this effect can be different in multiple tissues/conditions. In such cases, a batch-by-condition interaction effect should be considered in order to properly remove the unwanted variation caused by batches (Mead 1988).

**Unknown confounding factors**

Even if all known batch effects induced by sample processing have been explicitly corrected for, some variation arising from other untraceable factors will always remain. One example of a parameter difficult to trace is cellular state when studying the dynamics of gene expression during differentiation. When samples from a given cell differentiation stage are collected using certain markers (e.g. cell surface markers), the variation affecting those markers may have important consequences on QTL detection. Dissected tissue samples can contain slightly varying fractions of individual cell types, leading to cell-type specific gene clusters that vary in a correlated manner. If the fraction of a certain cell-type in the tissues happens to correlate with the genotype at a particular marker, all those cell-type specific genes will erroneously map to this marker. Therefore, it is suggested to perform genetical genomics experiments on samples from highly purified tissues or even at a single cell level (Gerrits et al. 2009); however, this would come with an inherent cost in terms of additional experimental and biological noise (Ozbudak et al. 2002).

Firstly, the presence of large non-genetic variation components in traits reduces the proportion of variance explained by genetic factors, thereby limiting the power for QTL detection.

Secondly, the large scale correlation induced by non-genetic variation can be dominant and cause coexpression methods to fail to capture the correlation stemming from genetic variation. Indeed, the impact of genetic variation on the correlation structure has been observed to be weak. For example, Ayroles et al. noted that the heritability of gene expression traits and connectivity are negatively correlated among a panel of D. melanogaster inbred lines (Ayroles et al. 2009), and suggested that hub genes are protected against genetic variation by purifying selection so that genetic variation does not have widespread effects within the coexpression network. Instead, it is likely that the global correlation structure observed in most experiments stems from variation in non-genetic factors such as sex, tissue compo-
sition, physiological state or experimental conditions. Eliminating such sources of correlation is at least as crucial to QTL based causal inference as it is to coexpression analyses, since correlated error terms are known to confuse causal inference methods (Li et al. 2010b).

Thirdly, several studies have suggested that most hotspots are likely to be spurious (Breitling et al. 2008a, de Koning and Haley 2005) and may be a mere statistical consequence of the correlation structure induced by hidden factors inherent to the data. If, by chance, one of those hidden factors correlates (even moderately) with the genotype distribution at one of the markers, a QTL hotspot containing most of these genes will be spuriously detected at this marker. In Box 8.2, several scenarios leading to such wrong interpretations of the data are detailed.

Since this confounding variation is usually not completely resolved by randomization of the experimental design (Churchill 2002) or through low-level normalization techniques (Yang et al. 2002), unsupervised methods to identify and control the variation associated with hidden factors have recently been proposed. For example, surrogate variable analysis (SVA) identifies the variables representing consistent expression signatures due to sources other than the factors of interest and these can be included in subsequent analyses as covariates to reduce dependency across genes, stabilize false discovery rate estimates, and improve the reproducibility of the analysis (Leek and Storey 2007). Kang et al. proposed inter-sample correlation emended (ICE) eQTL mapping which estimates the total correlation between samples and incorporates it into a linear mixed model as a variance component accounting for a random effect (Kang et al. 2008). These methods result in the removal of most hotspots and in a more sensitive detection of local QTL signals. From the simulations results by Kang et al., ICE is able to correct for a mixture of strong and moderate confounding effects while SVA is able to correct only for a number of strong confounding factors. However, as Kang et al. acknowledge, true regulatory hotspots may affect a large number of genes and cause inter-sample correlation. Therefore, when correcting for this inter-sample correlation, one risks removing true hotspots as well. Finally, Dubois et al. identified large scale effects via Principal Component Analysis (PCA) on a comprehensive collection of unrelated datasets spanning multiple tissues and conditions, and used these principal components (termed transcriptional components) to correct for variation in their own independent data, resulting in improved power to detect QTLs (Dubois et al. 2010).

Importantly, even some of the correlation induced by genetic variation may falsely suggest a pathway relationship. For example, in an inbred genetic panel such as an F2 or RILs, large fractions of the genome are under strong linkage disequilibrium as
few recombination events are present within the studied population. Consequently, all genes under local genetic control within those areas are linked to a common genotype and therefore are (somewhat) correlated but this correlation between neighboring genes does not reflect membership to a common functional group. It is possible to remove such linkage disequilibrium-induced correlation by working with expression data corrected for local genotypes (Heap et al. 2009).

8.5 Conclusions

The adaptation of old concepts from classical genetics and epidemiology to the new postgenomic fields is establishing itself as a major research area with the potential to elucidate the biological processes leading to complex phenotypes. As standard good practices are adopted by the community for the design, statistical analysis and biological interpretation of genetical genomics experiments, the trend of these genetic studies will be to go deeper (integrating more molecular levels (Ferrara et al. 2008, Fu et al. 2009, Johannes et al. 2008) and broader (larger sample sizes, combining genetic perturbation with other factors such as environmental factors) (Jansen et al. 2009, Li et al. 2008). The pervasive correlation structure stemming from (mainly poorly understood) physiological and technical factors within genomics datasets is appearing as the main challenge slowing down the path towards new discovery. Promising new approaches that tackle this confounding variation (Dubois et al. 2010, Kang et al. 2008, Leek and Storey 2007) are emerging and already proving to be beneficial as they improve the power to detect QTL while eliminating spurious findings. The application of these new approaches to network reconstruction (Chaibub Neto et al. 2008, Chen et al. 2007, Schadt et al. 2005, Zhang and Horvath 2005) promises to be accompanied by new breakthroughs by removing one of the major obstacles on the way towards reliable network inference (Li et al. 2010b).